中 文 信 息 学 报
第16卷 第6期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.16 No.6

一种基于上下文的中文信息检索查询扩展*

贺宏朝[1] 何丕廉[1] 高剑峰[2] 黄昌宁[2]

(1. 天津大学电子信息工程学院 天津 300072 2. 微软(亚洲)研究院 北京 100080)

摘要:在中文信息检索的研究和实践中,由于查询中所使用的词可能与文件集中使用的词不匹配而导致一些相关的文件不能被成功地检索出来,这是影响检索效果的一个很关键的问题。查询扩展可以在一定程度上解决这种词的不匹配现象,然而,实验表明,通常简单的查询扩展并不能稳定地提高中文信息检索的检索效果。本论文中提出并实现了一种基于上下文的查询扩展方法,可以根据查询的上下文对扩展词进行选择,是一种相对"智能"的查询扩展方法。在 TREC-9 中文信息检索测试集上进行的实验表明,相对于通常简单的查询扩展,基于上下文的查询扩展方法取得了具有统计意义提高的检索效果。

关键词:查询扩展;基于上下文;中文信息检索

中图分类号:TP391.3

Query Expansion Based on the Context in Chinese Information Retrieval

HE Hong-zhao[1] HE Pi-lian[1] GAO Jian-feng[2] HUANG Chang-ning[2]

(1. School of Electronic Information Engineer Tianjin University Tianjin 300072 2. Microsoft Research Asia Beijing 100080)

Abstract: Term mismatch between queries and documents is a fundamental problem in Chinese Information Retrieval (IR), which affects the effectiveness of retrieval results. Query expansion in IR can deal with this kindof problem in some degree. However, experiments show that the common query expansion in IR cannot get steady retrieval results. In this paper, we propose and realize query expansion based on the context, which can choose the expansion words according to the context of the query. Experiment results with TREC-9 show that query expansion based on the context is a smarter method. Compared with the results of common query expansion, query expansion based on the context can get statistically significant improvement.

Key words: query expansion; based on the context; Chinese information retrieval

一、引言

信息检索研究中,存在一个很关键的问题,由于用户所选择使用的词可能与文件集中所出现的词不匹配,导致检索效果降低甚至失败。比如,用户使用的词为"电脑",而文件集中出现的却都是"计算机",尽管,"电脑"与"计算机"描述的是完全相同的概念,但对于通常的信息检索系统来讲,"电脑"与"计算机"被认为是完全不同的两个检索单元。于是,由于这种词的不匹配现象导致一些相关的文件不能被成功的检索出来。

* **收稿日期**:2002-6-24
基金项目:天津市自然科学基金项目(993800111)
作者贺宏朝,男,1972 年生,博士研究生,主要研究方向为自然语言处理、中文信息检索、跨语言信息检索、本论文的大部分工作完成于作者在微软(亚洲)研究院访问期间. 何丕廉,男,1943 年生,教授,博士生导师,主要研究方向为人工智能、自然语言处理、计算机辅助教育.

对查询进行有利于检索的扩展,比如,对查询“电脑”进行扩展之后变为“电脑/计算机/电子计算机”,从而使查询中包含更多的相关信息,可以有效地解决大部分词的不匹配现象,达到提高检索效果的目的。

查询扩展的方法有很多,除了人工进行查询扩展的方法以外,其它的方法基本可以分为两大类:(1)对第一次检索的结果进行分析并从中选出更多的信息加入查询中;(2)利用某种资源对查询直接进行扩展。

对于第(1)类的查询扩展,最简单的方法就是“伪相关反馈”,就是假定在第一次检索的结果中,排在最前面的 n(n 为任意整数,通常可取 10,100 等)篇文件是与查询相关的,然后对这 n 篇文件进行统计,选择其中 m(m 为任意整数,通常可取 50,100 等)个词加入到查询中,并利用扩展后的查询重新进行检索。

第(2)类的查询扩展需要利用某种包含有词与词间相关信息的资源来进行,这种资源可以是人工生成的也可以是利用大规模语料通过统计的方法自动生成的。WordNet[1]就是一种人工生成的资源,它提供了英文单词之间的复杂关系,包括同义词、反义词、修饰词等词与词间的相关信息,Rila[2]利用 WordNet 中所提供的信息进行了英文信息检索查询扩展的研究,取得了相对积极的效果。

最近几届 TREC 会议的研究结果表明,使用上述第(1)类的查询扩展方法,通常可以较显著地提高信息检索的检索效果。但同时,也有研究表明,这些查询扩展方法的效果并不稳定,其效果强烈依赖于第一次检索的结果,如果对于某查询,第一次检索所得到的前 n 篇文件中,只有极少数的文件的确与该查询相关(比如,由于词的不匹配现象),那么在这种情况下,最可能的情况是将一些与该查询不相关的词作为扩展词加入到查询中,如此的查询扩展会使信息检索的效果迅速降低[3]。因此,通常先考虑进行第(2)类的查询扩展,在获得相对更可靠的检索结果之后,再进行第(1)类的查询扩展。

在本论文中,我们将只限于讨论中文信息检索中的第(2)类的查询扩展,研究利用中文的同义词资源进行中文信息检索查询扩展。不同于普通的查询扩展,我们提出并实现了一种基于上下文的查询扩展的方法,该方法可以根据查询的上下文对扩展词进行选择,是一种相对“智能”的查询扩展方法。在 TREC-9 中文信息检索测试集上进行的实验表明,相对于通常简单的查询扩展方法,基于上下文的查询扩展方法取得了具有统计意义提高的检索效果。

本论文的结构如下:第一部分概括地介绍了信息检索查询扩展;第二部分提出了一种基于上下文的查询扩展的方法;第三部分详细介绍了实验设计和结果;第四部分是结束语。

二、基于上下文的查询扩展

实际的文件中,对于同样的一词,当位于不同的上下文时,所描述的概念在内涵上通常是会有所差异的,有的差别甚至很大。比如“人们/发现/这个/地区/有/金子”与“小王/拥有/一颗/金子/般/的/心”的两句话中,虽然都出现了“金子”这个词,但其意义显然不同。如果利用同义词资源进行机械的扩展(比如在通常的查询扩展方法中),对于这样的同一个词,无论其是否具有相同的上下文,为其所选择的扩展词将是一样的。从而,这种与上下文相关的内涵上的差异势必会被掩盖掉,进而,最终影响到信息检索的交果。因此,当进行信息检索的查询扩展时,有必要结合词的上下文来选择扩展词。为此,我们提出了一种“基于上下文的查询扩展”方法,该方法在选择扩展词时,可以做到基于整个查询的主要内容而不是仅着眼于一个孤立的词,是一种相对更“智能”的查询扩展的方法。

在信息检索的研究中,我们可以使用统计方法(比如,词共现模型)得到词与词之间的相关性信息,记作 $SIM(x, y)$ (词 x 与词 y 之间的相关性)。通常,一条查询 q 会由若干个词组成,从而形成一个由词组成的集合 T_q。整个查询的主要内容是由 T_q 中的所有元素(词)来共同表现的。因此,定义词 x 与查询 q 的相关性为: $Cohension(x, q)$:

$$Cohesion(x, T_q) \log\left(\sum_{y \in T_q} SIM(x, y)\right) \quad (1)$$

利用式(1),我们就可以得到词 x 与整个查询 q 的相关性 $Cohension(x, q)$,而不是孤立的词与词之间的相关性 $SIM(x, y)$。这样,通过利用词与整个查询的相关性 $Cohension(x, q)$,就可以将与整个查询的主要内空相关的候选词选择为扩展词。于是,对于同样的一个词,如果其所在查询所表达的主要内容(上下文)不同,那么,所选择的扩展词也会是有区别的。也就是说,通过利用词与整个查询的相关性 $Cohension(x, q)$,我们做到了在考虑词的上下文的基础上,进行查询扩展。

本论文中,我们首先利用同义词资源为查询中的每一个词提供一个扩展词侯选集,然后,利用式(1),在这个扩展词侯选集中进行基于上下文的选择,最后将所选得扩展词加入到原查询中,进行信息检索。在 TREC-9 中文信息检索没试集上进行的实验表明,相对于通常简单的查询扩展方法,基于上下文的查询扩展方法取得了具有统计竟义提调换检索效果。

三、实验设计和结果

3.1 测试集

由美国技术和标准研究所(TNTST)主办的 TREC(Text REtrieval Conference),发起并资助了建立大规模信息检索标准测试集的资源建设工程,为信息检索的研究提供了科学的标准测试集。TREC-9 中文信息检索标准测试集中的文件集是由香港商报(1998 年 8 月 11 日到 1999 年 7 月 31 日),香港日报(1999 年 2 月 1 日到 1999 年 7 月 31 日)以及大公报(1998 年 10 月 21 日到 1999 年 3 月 4 日)中的文章收录组成,共有 127,938 篇文件,其查询共有 25 条,并由 TREC 组织人力完成了与各条查询相关文件的筛选工作。本论文中的实验是在 TREC-9 中文信息检索标准测试集上进行的。

3.2 信息检索系统

SMART 信息检索系统最早是由 Salton 于六十年代后其实现的基于向量空间检索模型(VSM)的信息检索系统[4],其最根本的目的是为信息检索研究提供一个研究框架,包括建立索引,检索和评价等基本功能。

作为一种非商业性的信息检索系统,SMART 信息检索系统的最新版本可以很方便的从网上获得(ftp: //ftp. cs. cornell. edu/ pub/ smart),目前,其版本为"version10/ 11/ 。经过了四十多年的研究和发展,最新版本的 SMART 信息检索系统已经是一个健壮的可处理大规模文件集(500MB)的信息检索系统。其检索效率也相对较理想,以建立索引为例,一小时可以完成 150MB 的文件处理(500MHz 的 Pentinum 处理器)。

SMART 信息检索系统还为用户提供了一定的灵活性,用户可以增加新功能或改变处理流程等。另外,由于其源代码公开,用户可以针对自己的需求对其进行修改,在本论文中,为便于进行中文信息检索的研究,使用的就是经过修改的 SMART 信息检索系统。

3.3 伪同义词资源

在实验中,我们利用 LDC 双语词典[5]建造了一个"伪同义词资源"。对于中文词 x,首先

通过查 LDC 中曲词典,得到词 x 的英文翻译,由于在词典中,通常一个中文词会对应若干条英文翻译,因此词 x 的英文翻译实际上组成了一个英文词集合 $\{e_1, e_2, e_3, \cdots e_m\}$。然后,再通过查 LDC 英中词典,将这个英文词集合中的所有英文词翻译成文,类似地,每一个英文词也对就了若干条中文翻译,因此,最终形成了一个由小集合组成的大集合。对最终的大集合进行适当的处理,就可以得到中文词 x 的"伪同义词资源"。之所以称之为"伪同义词资源",就是因为该资源并不是完全意义上的同义词资源,而是假定在不同语言中(比如中文和英文),概念在内涵和外延上具有一定程度的相似性,借助双语词典,通过在两种语言间的两次翻译而得到的。

3.4 带衰减因子的词共现模型

本论文中,为了能够更精确地描述词与词间的相关信息,我们使用带衰减因子的词共现模型来获得词 x 与词 y 之间的相关性 $SIM(x, y)$。

词共现模型是建立在这样一个基本假设的基础之上,如果在大规模语料(训练语料)中,两个词经常共同出现(共现)在同一窗口单元(如,文件中 5 个相邻的词组成的窗口单元)中,则认为这两个词在竟义上是相互关联的,而且共现的频率越高,其相互间的关联越紧密。基于这样的一个假设,通过对训练语料的统计,计算得到词与词之间的互信息(Mutual Information)。利用词共现模型可以对词与词之间的相关性进行量化的比较,正因如此,词共现模型更多地被用于跨语言信息检索对查询的翻译侯选词进行选择,与此相关的众多研究结果表明使用词共现模型取得了相对较理想的效果[6],这也证明词共现模型的基本假设具有一定的合理性。

通常定义词与词之间的互信息 $MI(x, y)$

$$MI(x, y) = P(x, y) \times \log\left(\frac{P(x, y)}{P(x) \times P(y)}\right) \tag{2}$$

其中,

$$P(x, y) = \frac{C(x, y)}{\sum_{x', y'} C(x', y')} \tag{3}$$

$$P(x) = \frac{C(x)}{\sum_{x'} C(x')} \tag{4}$$

式(3)中,$C(x, y)$ 是指在训练语料中,词 x 与词 y 在同一个窗口单元(比如,文件中 5 个相邻的词组成的窗口单元)中同时出现的频率。式(4)中,$C(x)$ 是指在训练语料中,词 x 出现的频率。

利用合适的训练语料,通过对 $C(x, y)$ 和 $C(x)$ 的统计,再利用式(2)进行计算就可以生成词与词之间的互信息资源。在基本的词共现模型中,直接使用词与词之间的互信息 $MI(x, y)$ 作为词与词相关性的度量。实验表明,如果直接利用这种自动生成的资源进行信息检索的查询扩展,其效果有可能并不令人满意,因此,有必要对这种基本的词共现模型进行改进。

首先是选择合适的窗口单元。在以往的研究中,人们选择了各种各样的窗口单元,如固定大小的窗口(比如,文件中 5 个相邻的词组成的窗口单元),文件中的自然段以及整篇文件等。我们认为,对于中文的文件而言,通常,一个句子表达一个相对较完整的意思,因此如果选择句子作为窗口单元,应该更能反映出意义上的相关性。因此,在我们的词共现模型中选择句子作为窗口单元。

我们注意到,如上所述的基本的词共现模型中,并没有考虑到对所考察的一对词之间的距离(词与词之间所出现的词的个数),无论一对词是紧邻的还是被很多词隔开,只要是在同一个

窗口单元中,都一致被认为是意义上相关的。事实上,即使在同一个窗口单元中,词之间的相关性也不会完全一致。为了使词共现模型能够反映出词之间的距离信息,我们在基本的词共现模型的基础上,进行了改进。

我们认为,在同一个窗口单元(句子)中,词与词之间的相关性是随着词之间的距离的增加而减少的。通过观察和尝试,我们假定,词与词之间的相关性随着词间距离指数衰减。因此,在式(2)的基础上,我们加入了一项反映词间距离信息的衰减项,如式(5)定义了词与词之间的相关性 $SIM(x,y)$,形成了改进的词共现模型,我们称它为带衰减因子的词共现模型。我们利用这样的带衰减因子的词共现模型进行了一系列有关于中文信息检索查询扩展的研究,取得了令人满意的结果。研究结果证明了我们的词与词之间的相关性是随着词间距离指数衰减的假设是合理的。

$$SIM(x,y) = MI(x,y) \times e^{-\alpha \times (D(x,y)-1)} \quad (5)$$

其中,$D(x,y)$是指在所有窗口单元(句子)中,词 x 与词 y 之间的平均距离。

α是常数,表示词之间相关性随词间距离进行衰减的剧烈程度,可以利用训练集通过实验确定。为此,我们选择 TREC - 9 的文件集作为训练集,利用 TREC - 9 标准测试集,使用 SMART 信息检索系统(文件和查询均选择 ltc 权值),进行了一系列的实验,以获得 α 的最佳取值。为了方便处理,对于查询中的每一个词,我们只选择了一个最佳的候选词作为扩展词。并以 0.2 为步长,分别设定 α 为 0 到 2.0 之间的 11 个值,如图 1 示出了在选择不同的 α 时,进行基于上下文的查询扩展后的检索结果。

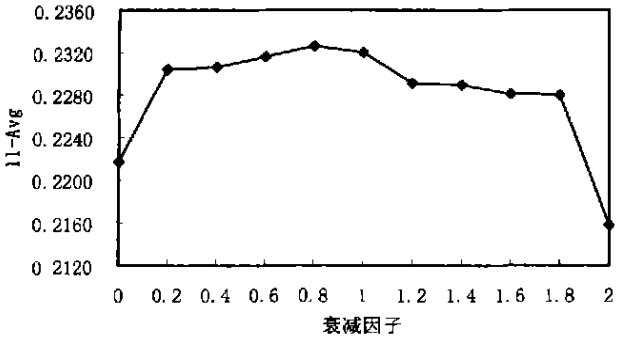

图 1　衰减因子对基于上下文的查询扩展检索结果的影响

如图 1 所示,基于上下文的查询扩展的检索结果随着衰减因子 α 的变化而发生变化,其中的 11-Avg 是指,在 11 个召回率点上(0,0.1,0.2,...,1.0),25 条查询所对应的精度的平均值的平均。

如果选取 $\alpha = 0$,由式(5)可知,此时的词共现模型实际上已经退化为基本的词共现模型(式 2)。很明显,当在合适的取值范围($0 < \alpha < 2$)内选择 α 时,由基于上下文的查询扩展的检索结果可知,带有衰减因子的词共现模型的检索结果比基本的词共现模型的结果都要好。而且,当选取 $\alpha = 0.8$ 时,取得了最好的检索结果,其 11-Avg 值为 23.26 %,相对于基本的词共现模型($\alpha = 0$)的结果,其 11-Avg 值具有 5 %的提升。因此,在以后的实验中,我们选定衰减因子 $\alpha = 0.8$。

我们的结果也证明了,带有衰减因子的词共现模型的合理性,与基本的词共现模型相对比,带有衰减因子的词共现模型更能有效地反映出词与词之间不同紧密程度的相关性。通过利用合适的训练集(在本文中,我们使用 TREC-9 文件集),应用上述带衰减因子的词共现模型,可以自动生成可用于中文信息检索查询扩展的包含有词与词相关性的资源。

3.5　实验结果

如前所述,我们利用 TREC-9 中文信息检索标准测试集,使用 SMART 信息检索系统,利用伪同义词资源为查询提供扩展词侯选集,使用带衰减因子的词共现模型获得词与词之间的相关性,进行了两组实验,分别对应于:(1)简单的查询扩展方法,(2)基于上下文的查询扩展方

法。其中,(1)是指普通的查询扩展,即在扩展词侯选集中选择与查询中该词最相关的词(不考虑查询中其它的词)作为扩展词。(2)为前述的基于上下文的查询扩展方法,其扩展词的选择是根据词与整个查询的主要内容的相关程度来进行,因此,是一种基于上下文的查询扩展方法。

图 2 示出了检索结果。为了方便处理,对于查询中的每一个词,两种查询扩展方法都只选择了一个最佳的候选词作为扩展词。其中的平均精度是指,在相应的召回率点,25 条查询所对应的精度的平均值。

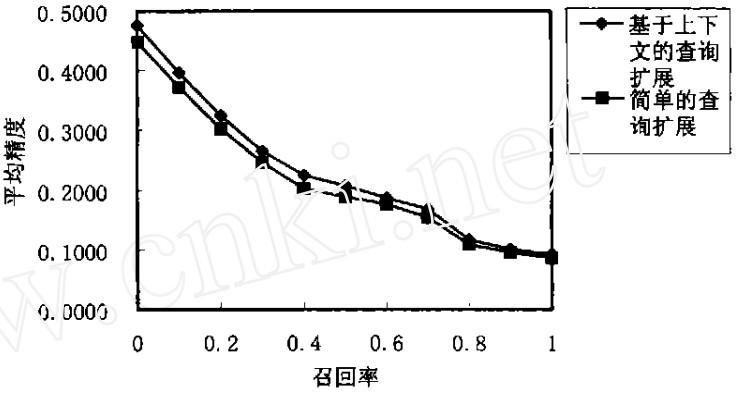

图 2　简单的查询扩展方法和基于上下文的查询扩展方法检索结果

我们的结果也证明了,带有衰减因子的词共现模型的合理性,与基本的词共现模型相对比,带有衰减因子的词共现模型更能有效地反映出词与词之间不同紧密程度的相关性。

3.6　分析和讨论

如图 2,基于上下文的查询扩展方法的检索结果明显好于简单的查询扩展方法的检索结果,其在 11 点召回率上平均精度平均提高了 12 %。

由于,平均精度和召回率曲线(如图 2)反映的是 25 条查询的平均结果,因此,我们对这两种方法所得到的每一条查询的检索结果进行的统计意义评价,以确定这种提高是一种随机的提高还是具有统计意义的显著提高。文献[7]中详细地讨论了信息检索中统计意义的评价方法,通常认为两组数据的 p-value 值小于 0.05 就可以认为是一种统计意义上的显著不同。利用我们的实验结果计算得到相应的 p-value 值为 0.02(小于 0.05),因此,可以断定相对于简单的查询扩展方法的检索结果,基于上下文的查询扩展方法取得了具有统计意义的显著提高的检索结果,从而也证明了基于上下文的查询扩展方法的确是一种更"智能"的扩展方法。

四、结束语

在信息检索的研究和实践中,通常利用查询扩展的方法来解决由于查询与文件集中词的不匹配现象导致检索效果降低的问题。本论文提出并实现了一种基于上下文的查询扩展方法,不同于通常的查询扩展方法,该方法可以根据查询的上下文对扩展词进行选择,是一种相对"智能"的查询扩展方法。在 TREC-9 中文信息检索测试集上进行的实验表明,相对于通常简单的查询扩展,基于上下文的查询扩展方法取得了具有统计意义提高的检索效果。

参 考 文 献

[1] Miller G A, et al. Introduction to WordNet: an on-line lexical database, International Journal of Lexicography, 1990, 3(4): 235 - 312

[2] Rila Mandala, Takenobu Tokunaga, Hozumi Tanaka, Combining multiple evidence from different types of thesaurus for query expansion, SIGIR, 1999: 191 - 197

(下转第 45 页)

gen ,Groningen ,The Netherlands ,1989

[4] Willems ,M. , *Chemistry of language : a graph theoretical study of linguistic semantics* , Ph. D. Thesis ,University of Twente ,Enschede ,The Netherlands ,ISBN 90 - 9005672 - 9 ,1993

[5] Whitehead ,A. N. and B. Russell , *Principia Mathematica* , Cambridge University Press ,Cambridge ,1925 ,2nd edition

[6] Huang ,K. ,Introduction to Expert Systems ,Southeastern University Press ,1988

[7] 刘小冬 ,李学良 ,张蕾. 知识图综述. 工程数学学报 ,2000(17)增 :33 - 40

[8] 新汉英词典. 北京 :外语教学与研究出版社 ,1988

(上接第 37 页)

[3] Voorhees E M ,Harman D K ,The sixth Test REtrieval Conference (TREC-6) , Gaithersburg ,NIST ,1998

[4] Salton G ,The SMART retrieval system-experiments in automatic document processing ,Prentice Hall ,1971 :115 - 411

[5] http : //morph. ldc. upenn. edu/ Projects/ Chinese

[6] Gao J F ,Nie J Y ,Zhang J ,et al ,Improving query translation for CLIR using statistical models ,ACM SIGIR '01 , New Orleans ,2001 :96 - 104

[7] David Hull ,Using statistical testing in the evaluation of retrieval performance , In Proc. of the 16th ACM/ SIGIR Conference ,1993 :329 - 338

会议消息

中国中文信息学会与国际中文计算机学会于 2003 年 8 月 3 日 —6 日在沈阳市召开“第二十届东方语言计算机处理国际学术会议”,会议由东北大学承办。

2003 年 8 月 8 日 —11 日中国中文信息学会与兄弟学会在哈尔滨市召开“全国第六届计算语言学联合学术会议(JSCL —2003)”。会议由哈尔滨工业大学承办。

上述两次会议的征文内容如下:

1) 计算语言学的理论研究;

2) 汉语的词汇、句法和语义;

3) 语料库建设、语料加工技术及基于语料库的语言分析技术;

4) 汉语的文本分析与生成;

5) 机器翻译系统、技术及评测方法;

6) 文本智能检索、文本自动分类、文本过滤及自动文摘;

7) 汉语语音识别与语音合成;

8) 智能型汉字输入方法;

9) 其它。

征文截稿日期等重要信息另行通知。